Presentation on

Beyond Benign Overfitting in Nadaraya-Watson Interpolators

Daniel Barzilai* Guy Kornowski* Ohad Shamir
Weizmann Institute of Science
{daniel.barzilai,guy.kornowski,ohad.shamir}@weizmann.ac.il

October 22, 2025

by Anay Mehrotra

Nadaraya–Watson (NW) estimator

Nadaraya–Watson (NW) estimator [Nadaraya, 1964; Watson, 1964]

$$f_{\beta}(x; S)$$

Hyperparameter *Training data*

Nadaraya–Watson (NW) estimator

Nadaraya–Watson (NW) estimator [Nadaraya, 1964; Watson, 1964]

$$f_\beta(x; S) := \begin{cases} \text{sign}\left(\sum_i \frac{y_i}{\|x - x_i\|^\beta}\right) & \text{if } x \notin S\,, \\ y_i & \text{if } x \in S \text{ and } x = x_i\,. \end{cases}$$

Hyperparameter *Training data*

- ➔ Nearest-neighbour based classification rule
- ➔ Displays benign overfitting (details next) [Devroye, Györfi, and Krzyżak, 1998]

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Testing error: Error is evaluated on *clean data*

$$\mathrm{Err}(f(S_p)) := \Pr_{x \sim \mathcal{D}}[f(x; S_p) \neq f^{\star}(x)]$$

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Testing error: Error is evaluated on *clean data*

$$\mathrm{Err}(f(S_p)) := \Pr_{x\sim\mathcal{D}}[f(x; S_p) \neq f^{\star}(x)]$$

Theorem **[Devroye, Györfi, and Krzyżak, 1998]** If D has d-dimensional support. Then, the NW-estimator with $\beta = d$, despite noise $p \in [0, 1/2)$ achieves:

$$\mathrm{Err}(f; S_p) \to 0 \qquad \text{as} \qquad |S_p| \to \infty$$

Benign Overfitting in Classification

Theorem [Devroye, Györfi, and Krzyżak, 1998] If D has d-dimensional support. Then, the NW-estimator with $\beta = d$, despite noise $p \in [0, 1/2)$ achieves:

$$\mathrm{Err}(f_d; S_p) \to 0 \qquad \text{as} \qquad |S_p| \to \infty$$

➔ Even though *f fits the noise in the data exactly,* it still *generalizes to clean data*

➔ Benign over-fitting is *not an entirely new observation!*

➔ Similar estimators also analyzed for, e.g., *regression*. *Do they benignly overfit too?*

Benign Overfitting in Classification *Continued*

Theorem [Barzilai, Kornowski, Shamir, NeurIPS'25] If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $|S_p| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$

Benign Overfitting in Classification *Continued*

Theorem **[Barzilai, Kornowski, Shamir, NeurIPS'25]** If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $|S_p| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$
- If $\beta = d$, then $\mathrm{Err}(f_d; S_p) \to 0$

Benign Overfitting in Classification *Continued*

Theorem **[Barzilai, Kornowski, Shamir, NeurIPS'25]** If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $|S_p| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$
- If $\beta = d$, then $\mathrm{Err}(f_d; S_p) \to 0$
- If $\beta > d$, then $\mathrm{Err}(f_\beta; S_p) \to [p^{O(1)}, O(p)]$

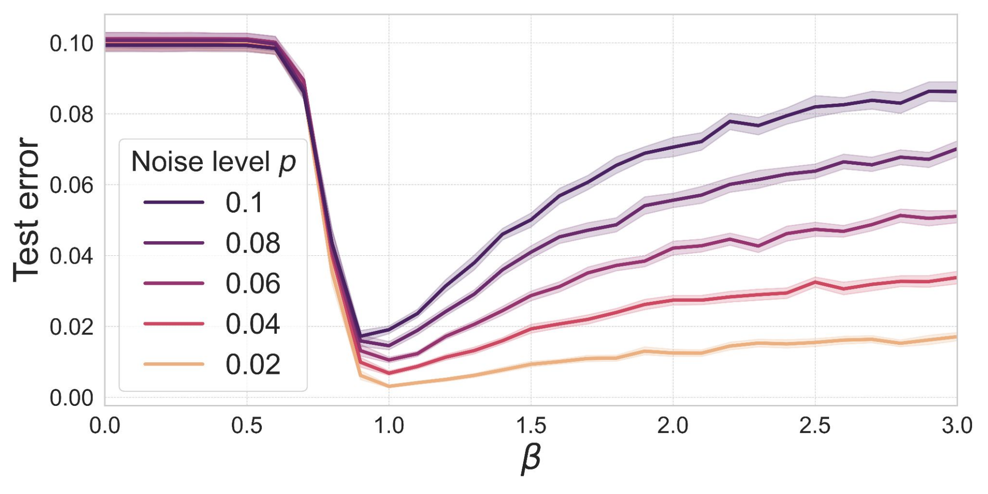
➔ Benign overfitting for the NW estimator is *fragile*…

➔ Right hyperparameter choice depends on the *ambient data dimension*

Empirical Results: 1-Dimensional Data

$\mathcal{D} := \text{Uniform}[0, 1]$

$f^\star(x) := \mathbb{1}\{x \in [0, 1/4]\}$

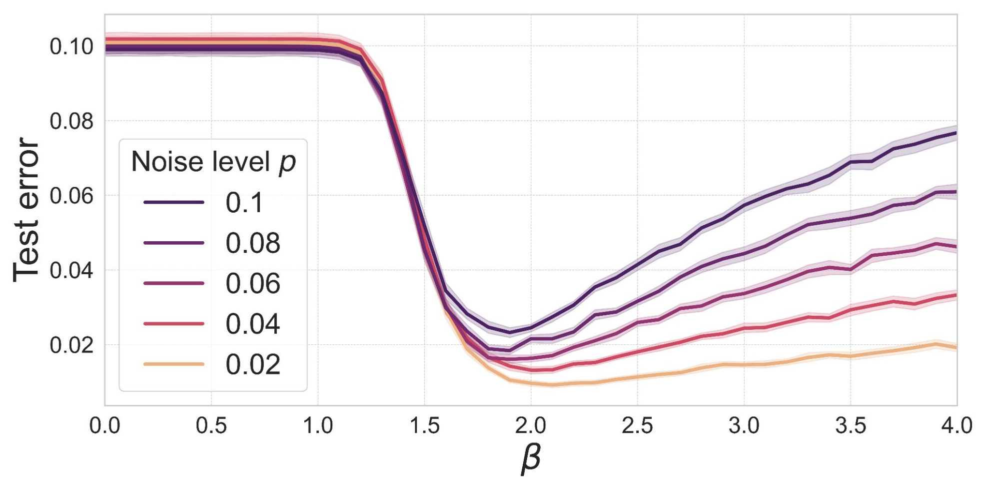


Empirical Results: 2-Dimensional Data

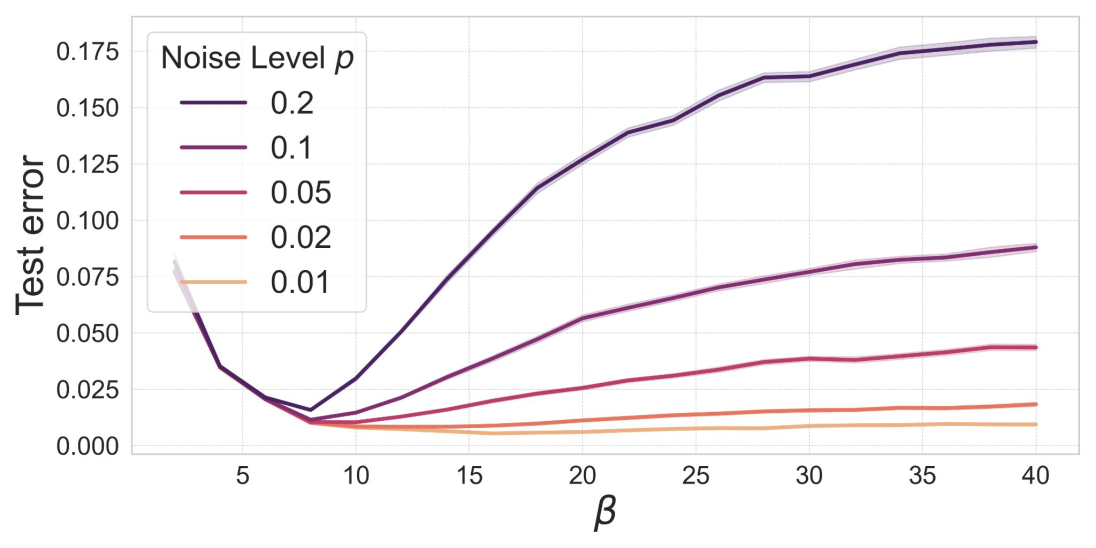
$$A := \left\{ x = (x_1, x_2, x_3) \in \mathbb{S}^2 \;\middle|\; x_3 > \frac{\sqrt{3}}{2} \right\},$$

$$\mathcal{D} = \frac{1}{10} \cdot \mathrm{Unif}(A) + \frac{9}{10} \cdot \mathrm{Unif}\left(\mathbb{S}^2 \setminus A\right)$$

$$f^\star(x) := \mathbb{1}\{x \notin A\}$$



Empirical Results: MNIST



➔ [Pope et al., ICLR’21] estimated MNIST’s intrinsic dimension to be in [8, 15]!